DOCUMENT RESUME

ED 430 001 TM 029 711

AUTHOR Roberts, James S.; Donoghue, John R.; Laughlin, James E.
TITLE Estimability of Parameters in the Generalized Graded Unfolding Model.
PUB DATE 1999-04-00
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); *Item Response Theory; *Maximum Likelihood Statistics; *Sample Size; Simulation
IDENTIFIERS *Generalized Graded Unfolding Model; Item Parameters; *Unfolding Technique

ABSTRACT

The generalized graded unfolding model (GGUM) (J. Roberts, J. Donoghue, and J. Laughlin, 1998) is an item response theory model designed to analyze binary or graded responses that are based on a proximity relation. The purpose of this study was to assess conditions under which item parameter estimation accuracy increases or decreases, with special attention paid to the influence that a given item parameter value has on the estimation of another item parameter. This assessment was based on a recovery simulation in which the effects of sample size, item location, degree of item discrimination, and extremity of subjective category thresholds were varied. Results indicate that with 750 or more respondents, sample size has negligible effect on all but the estimation of subjective response category thresholds. The true extremity of both item location and item discrimination did affect the estimation of these parameters themselves, and also affected the estimation of other item parameters in the model. However, these effects were modest and had little impact on the estimation of the corresponding item response functions. These results suggest that marginal maximum likelihood estimates of item parameters will provide accurate results across a variety of item parameter configurations when the sample size is at or above the recommended levels. (Contains 1 table, 3 figures, and 14 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Estimability of Parameters in the Generalized Graded Unfolding Model

ED 430 001

James S. Roberts
Medical University of South Carolina

John R. Donoghue
Educational Testing Service

James E. Laughlin
University of South Carolina

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

James Roberts

Presented at the annual meeting of the American Educational Research Association, April 22, 1999, Montreal, Canada.

Visit our web site devoted to item response theory models for unfolding:
www.musc.edu/cdap/roberts

TM029711

Correspondence should be directed to Dr. James S. Roberts, Medical University of South Carolina, Institute of Psychiatry - 4 North, 67 President Street, P.O. Box 250861, Charleston, SC 29425; E-mail: roberjam@musc.edu.

BEST COPY AVAILABLE

Abstract

The generalized graded unfolding model (GGUM; Roberts, Donoghue & Laughlin, 1998, 1999) is an item response theory model designed to analyze binary or graded responses that are based on a proximity relation (Coombs, 1964). A typical application of the GGUM is in measurement situations where respondents are asked to indicate their level of agreement with a series of statements that span a bipolar attitude continuum (e.g., Thurstone or Likert attitude measurement scales).

Roberts, Donoghue, and Laughlin (1998) have shown that when the data conform to the GGUM, then accurate item parameter estimates can be obtained with a marginal maximum likelihood procedure when the sample size is approximately 750 or more. Similarly, Roberts et al. have demonstrated that accurate expected *a posteriori* estimates of person parameters can be obtained with approximately 20 items with 6 response categories per item. Although the minimum data demands associated with these estimation procedures have been investigated, other characteristics about the robustness of parameter estimation accuracy remain unanswered.

The purpose of this study was to assess conditions under which item parameter estimation accuracy increases or decreases, with special attention paid to the influence that a given item parameter value has on the estimation of another item parameter. This assessment was based on a recovery simulation in which the effects of sample size, item location, degree of item discrimination and extremity of subjective category thresholds were varied. The results indicated that with 750 or more respondents, sample size had negligible effects on all but the estimation of subjective response category thresholds. Additionally, the true extremity of both item location and item discrimination did affect the estimation of these parameters themselves, and also affected the estimation of other item parameters in the model. However, these effects were modest and had little impact on the estimation of the corresponding item response function. These results suggest that marginal maximum likelihood estimates of item parameters will provide accurate results across a variety of item parameter configurations when the sample size is at or above the recommended levels.

Educational researchers typically use self-report questionnaires to assess attitudes toward or preferences for a variety of stimuli (e.g., attitude toward mathematics, preference for alternative types of instruction, etc.). Such questionnaires often contain a graded disagree-agree response format to gauge the level of individual agreement to a series of statements that range in content from negative, to neutral, to positive opinions. Several researchers (Andrich, 1996; Roberts, 1995; Roberts & Laughlin, 1996a, 1996b; Roberts, Laughlin & Wedell, 1999; van Schuur & Kiers, 1994) have argued that graded disagree-agree responses are generally more consistent with an unfolding model of the response process rather than the more popular cumulative model. Unfolding models are proximity models which imply that higher item scores, indicative of stronger levels of agreement, are more probable as the distance between an individual and an item on the underlying latent continuum decreases (Coombs, 1964).

Roberts and colleagues (Roberts, 1995; Roberts, Donoghue & Laughlin, 1998, 1999; Roberts & Laughlin, 1996a, 1996b) have developed a family of item response theory models that implement an unfolding response mechanism. The most general of these models is called the Generalized Graded Unfolding Model (GGUM). The GGUM defines the probability that the *jth* respondent will choose the *kth* response category to the *ith* item as:

$$Pr[Z_i = z \mid \theta_j] = \frac{\exp\left(\alpha_i [z(\theta_j - \delta_i) - \sum_{k=0}^{z} \tau_{ik}]\right) + \exp\left(\alpha_i [(M-z)(\theta_j - \delta_i) - \sum_{k=0}^{z} \tau_{ik}]\right)}{\sum_{w=0}^{C}\left[\exp\left(\alpha_i [w(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_{ik}]\right) + \exp\left(\alpha_i [(M-w)(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_{ik}]\right)\right]} \quad (1)$$

for $z = 0, 1, \ldots, C$; where θ_j is the location of the *jth* individual on the latent continuum; δ_i is the location of the *ith* item on the latent continuum; α_i is the discrimination parameter for the *ith* item; τ_{ik} is the *kth* subjective response category threshold for the *ith* item; C is the number of observable response categories minus 1; M is equal to $2*C+1$. Note that in the context of attitude measurement, θ_j is an index of the *jth* individual's attitude, and δ_i is an indicator of the *ith* item's affective content.

The GGUM leads to single-peaked, bell-shaped response functions that imply higher levels of agreement to the extent that the individual and the item are close to each other on the latent continuum. The GGUM is more general than other unfolding item response theory models in that it allows items to vary in their discrimination capabilities via α_i, and it allows subjects to utilize the response scale differently for each item via τ_{ik}. Figure 1 illustrates how the α_i and τ_{ik} parameters affect the item response function under the GGUM. In the two upper panels of Figure 1, the value of α_i is relatively low (i.e., .59) whereas it is relatively high (i.e., 1.63) in the two lower panels. Similarly, the distance between successive τ_{ik} values is lower (i.e., .25) for the two panels on the left-hand side of Figure 1, whereas this distance is higher (i.e., .57) for the two panels shown on the right-hand side of the figure. As shown in Figure 1, the item response function under the GGUM has a larger maximum value and becomes more peaked when the value of

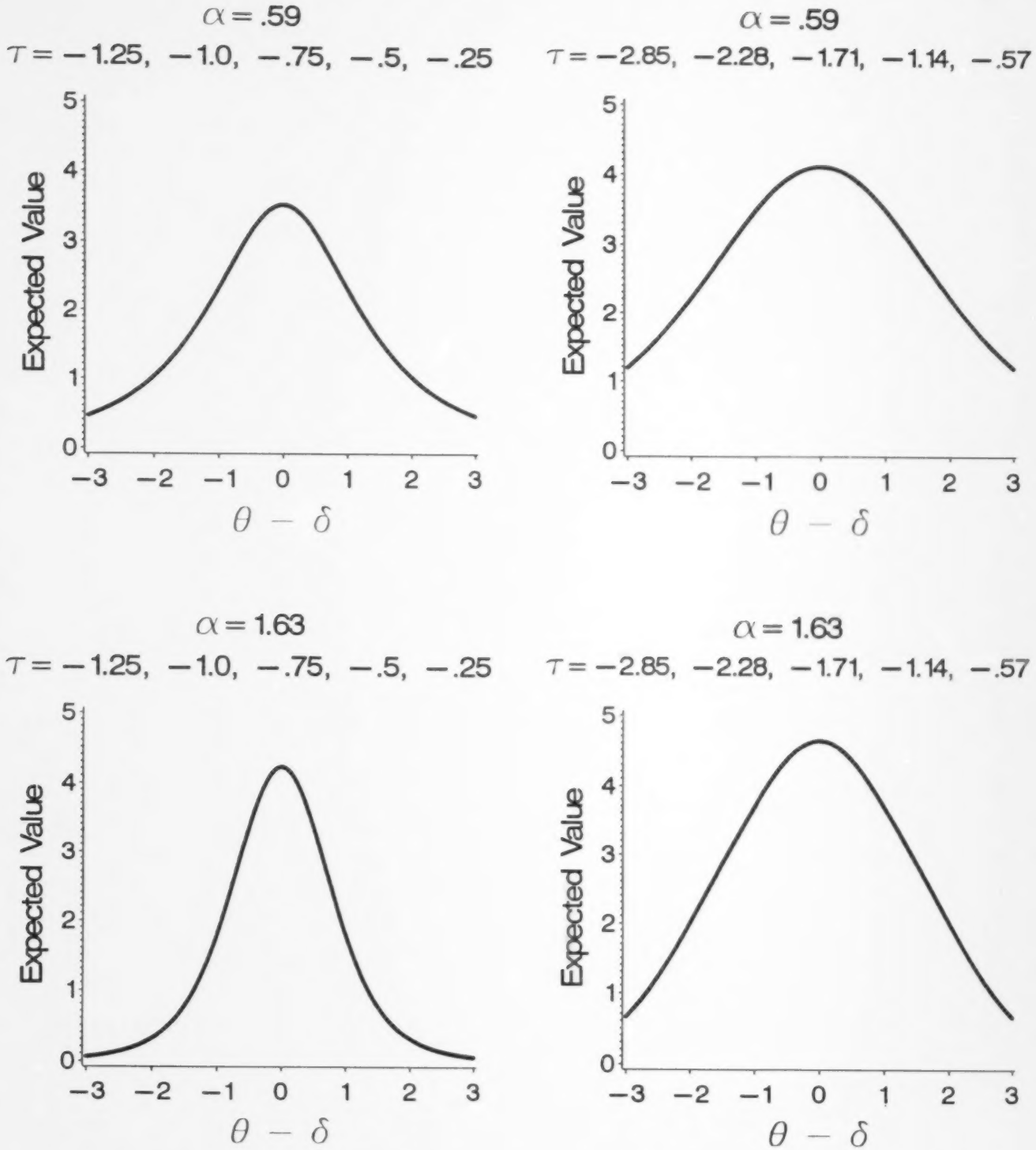

Figure 1. Item response functions for four hypothetical items under the GGUM. Items vary with regard to α_i (.59 or 1.63) and the distance between successive τ_{ik} values (.25 or .57). Each hypothetical item has 6 possible response categories that range from 0 to 5.

the item discrimination parameter increases. When the distance between successive τ_{ik} values increases, the item response function also achieves a larger maximum value but simultaneously becomes more diffuse. Thus, the values of α_i and τ_{ik} have distinctively different effects on the shape of the item response function under the GGUM.

Roberts et al. (1998) have shown that when data conform to the GGUM, item parameters can be accurately estimated using a marginal maximum likelihood (MML) technique (Bock & Aitkin, 1981; Muraki, 1992) with samples of 750 or more respondents. Additionally, person locations (i.e., attitudes or ideal points) can be accurately estimated with an expected *a posteriori* technique (Bock & Mislevy, 1982) when there are approximately 20 items with 6 graded disagree-agree response categories per item. Although these minimum data demands required to estimate parameters of the GGUM have been demonstrated, further questions about parameter estimability remain. Specifically, there is currently no information on the degree to which estimates of one parameter are affected by the values of other parameters in the model. One could speculate that the estimability of a given type of item parameter might degrade when values of the other parameters in the model get too large or too small. For example, it is important to know whether estimates of two identical item locations will be stable if the discrimination parameters associated with the two items vary widely. Similar questions can be asked with regard to estimates of each item parameter.

The primary aim of this study was to determine the accuracy of item parameter estimates as the values of remaining item parameters were varied in a systematic fashion. The utility of the GGUM model rests on the ability to accurately estimate model parameters, and thus, this issue is important to the advancement of this class of models in educational research.

Method

A parameter recovery simulation was used to study the effects of the following variables on the recovery of item parameters:

1) sample size (750 or 2000)

2) item location (-2, -1, 0, +1, +2)

3) discrimination (low=.59 ; high=1.63)

4) distance between successive subjective response category thresholds (a.k.a. interthreshold distance; low=.25; high=.57)

The values used for high and low discrimination and interthreshold distance corresponded to those observed in real data, whereas the range of item locations was chosen to be similar to past simulation research on item parameter recovery. The levels of each item parameter variable were crossed to produce a set of 20 items, and these same 20 items were studied under the 2 sample

size conditions. Responses to the 20 items were generated according to the GGUM, and item parameters were subsequently estimated using the MML technique. The true θ_j parameters used in the simulation were drawn from a N(0,1) distribution and were integrated out of the likelihood equation using a standard normal prior distribution (Bock & Aitkin, 1981; Roberts et al., 1998). Integration was accomplished with numerical quadrature using 30 equally spaced quadrature points ranging from -4 to +4. Convergence of parameter estimates was operationally defined as a change of less than .001 in any item parameter from one iteration of the MML algorithm to the next. This process of generating data and estimating parameters was independently replicated 30 times in each sample size condition using the same true values of person parameter estimates.

The recovery simulation formed a 2 (sample size) x 5 (item location) x 2 (item discrimination) x 2 (subjective response category thresholds) factorial design. Main and interaction effects in this design were analyzed using an analysis of variance (ANOVA) model. Four dependent measures were analyzed with the ANOVA model. The first three measures were the root mean squared error (RMSE) between the true and estimated values of the three types of item parameters in the GGUM (i.e., the RMSE for estimates of δ_i, α_i, and τ_{ik}) The fourth dependent variable was the average absolute deviation (AAD) between estimated and actual response functions over the interval of $\theta_j = (-3.0, +3.0)$.

The RMSE for the subjective response category threshold parameters associated with the *ith* item was calculated as:

$$RMSE = \sqrt{\frac{\sum_{k=1}^{C} (\hat{\tau}_{ik} - \tau_{ik})^2}{C}}, \tag{2}$$

where $\hat{\tau}_{ik}$ and τ_{ik} denote the estimated and true values for the *kth* subjective response category threshold, respectively. The RMSE for the *ith* item location was calculated as:

$$RMSE = \sqrt{(\hat{\delta}_i - \delta_i)^2}, \tag{3}$$

and that for the *ith* discrimination was computed as:

$$RMSE = \sqrt{(\hat{\alpha}_i - \alpha_i)^2}. \tag{4}$$

Therefore, the RMSE for the δ_i and α_i parameters was simply equal to the absolute deviation between estimated and true parameter values. Note that there was an RMSE value corresponding

to each of the three types of parameters associated with a given item in a particular replication.

The fourth dependent variable, the AAD, was analyzed in an analogous ANOVA model. Some researchers (Hulin, Lissak & Drasgow, 1982; Linn, Levine, Hastings & Wardrop, 1981) have suggested that this measure is more relevant than those which examine the accuracy of a single type of item parameter. This view stems from the fact that inaccuracies in estimates may cancel out across alternative item parameters, and thus, the estimated item response function might still be quite precise in cases where the accuracy of specific item parameters is questionable.

Due to the large number of observations used in these ANOVAs (N=1200), there was a substantial degree of power to detect both main effects and interactions in the 4 dependent measures. Therefore, only those effects that were statistically significant ($p<.0125$) and had a reasonable effect size ($\eta^2 \geq .05$) were interpreted. The type I error rate was set using a Bonferroni correction to control for the fact that there were 4 dependent measures examined using the same univariate ANOVA model ($\alpha = .05/4 = .0125$). The η^2 index was defined as the sum of squares for a given effect divided by the total sum of squares for the dependent variable.

Results

Overall Prediction of Individual Parameters and AAD

The average RMSE for δ_i, α_i, τ_{ik} parameters was equal to .07, .06, and .16, respectively. This represented 4.8%, 11.2% and 20.3% of the standard deviations of corresponding true item parameters. Thus, the τ_{ik} parameters were hardest to accurately estimate, although all item parameters were estimated to a reasonably accurate degree. The average AAD value was equal to .064, and thus, the average degree of inaccuracy associated with a given item response function was quite small relative to the range of the response scale (e.g., from 0 to 5).

ANOVA Results

The η^2 values and the significance level corresponding to each ANOVA effect are shown in Table 1 for the 4 dependent measures of interest. Only the main effects of sample size, true δ_i, and true α_i had effects that met interpretability criteria for any of the dependent measures. Given the high power associated with this design, the η^2 feature of the interpretability criterion was the dominant feature (i.e., all effects with $\eta^2 \geq .05$ were statistically significant at the $p < .0125$ level).

Effects of Sample Size

As expected, all item parameter estimates were more accurately estimated with the larger sample size of 2000 rather than the recommended sample size of 750. However, the increased

Table 1. η^2 values from ANOVAs on estimation accuracy indices.

Source Effect	RMSE $\hat{\delta}$	RMSE $\hat{\alpha}$	RMSE $\hat{\tau}$	AAD
N	**.04**	**.04**	**.12**	**.19**
δ	**.11**	.01	**.05**	**.02**
N x δ	**.03**	**.01**	**.02**	.01
α	**.05**	**.16**	**.24**	**.05**
N x α	**.01**	**.01**	**.04**	**.01**
δ x α	**.03**	.01	**.01**	.00
N x δ x α	**.01**	**.01**	**.01**	.00
τ	.00	**.01**	.00	.00
N x τ	.00	.00	.00	.00
δ x τ	.00	**.01**	**.01**	**.04**
N x δ x τ	.00	**.01**	.00	.01
α x τ	.00	.00	**.00**	.00
N x α x τ	.00	.00	**.00**	.00
δ x α x τ	.00	**.01**	.00	.01
N x δ x α x τ	.00	**.02**	.00	.01

Statistically significant effects are given in bold type. The type I error rate was equal to .0125. An effect was deemed worthy of interpretation when it was both statistically significant and accounted for at least 5% of the variation in a given dependent measure. N=sample size.

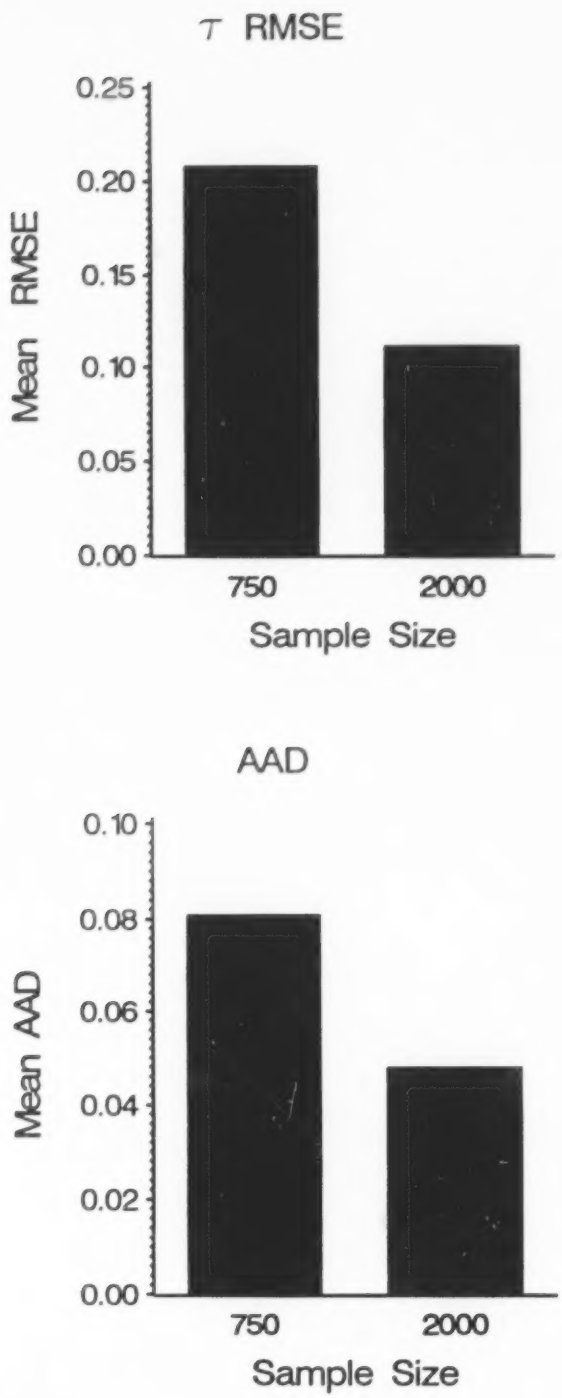

Figure 2. The effect of sample size on average RMSE values for τ_a (top panel) and mean AAD indices (bottom panel).

accuracy reached interpretable levels only for the RMSE of τ_{ik}. As shown in the top panel of Figure 2, the average RMSE for this parameter was equal to .11 when N=2000 and grew to .21 when N=750. These RMSE values represented 14.1% and 26.4% of the standard deviation of true τ_{ik} parameters, respectively.

The sample size also had an interpretable effect on the mean AAD measures as shown in the lower panel of Figure 2. However, the difference between AAD measures obtained in the two sample size conditions was only .03. Thus, this difference seemed minor in the context of the 6-point response scale.

Effects of True α_i

Figure 3 illustrates the effects of true α_i on the accuracy of item parameter estimates and the estimated item response function. The accuracy of all parameter estimates was affected by the degree of item discrimination. Both δ_i and τ_{ik} were estimated more poorly when the true α_i was low rather than high (RMSE δ_i : .10 versus .04; RMSE τ_{ik} : .23 versus .09). This "cross parameter effect" suggested that δ_i and τ_{ik} were more difficult to estimate when the true item response function was more diffuse. Interestingly, α_i itself was more accurately estimated when it was low rather than high (RMSE α_i : .03 vs .08). This finding constituted a "same parameter effect." Thus, item discriminations were harder to estimate when they were relatively high, yet item locations and subjective category thresholds were more precisely estimated when item discriminations were high.

The mean AAD index also varied systematically as a function of the true value of α_i. Specifically, the mean AAD index was slightly higher for the low discrimination condition relative to the high discrimination condition (e.g., .07 versus .06). Although this effect suggested more precise estimation of item response functions in the high discrimination condition, the difference is inconsequential when interpreted within the context of the 6-point response scale.

Effects of True δ_i

As shown in Figure 4, the ability to accurately estimate both δ_i and τ_{ik} parameters was affected by the extremity of δ_i. Specifically, the average RMSE of δ_i estimates increased as the absolute value of the corresponding true δ_i increased (i.e., a same parameter effect). The range of this increase was equal to .10. Similarly, the inaccuracy of τ_{ik} estimates also increased as the extremity of true δ_i grew larger (i.e., a cross parameter effect). The average RMSE for τ_{ik} estimates increased from .13 to .20 as δ_i became more extreme. Both the same parameter and cross parameter effects associated with true δ_i were presumably due to the fewer number of subjects (i.e., the smaller amount of information) available at the more extreme positions on the latent continuum.

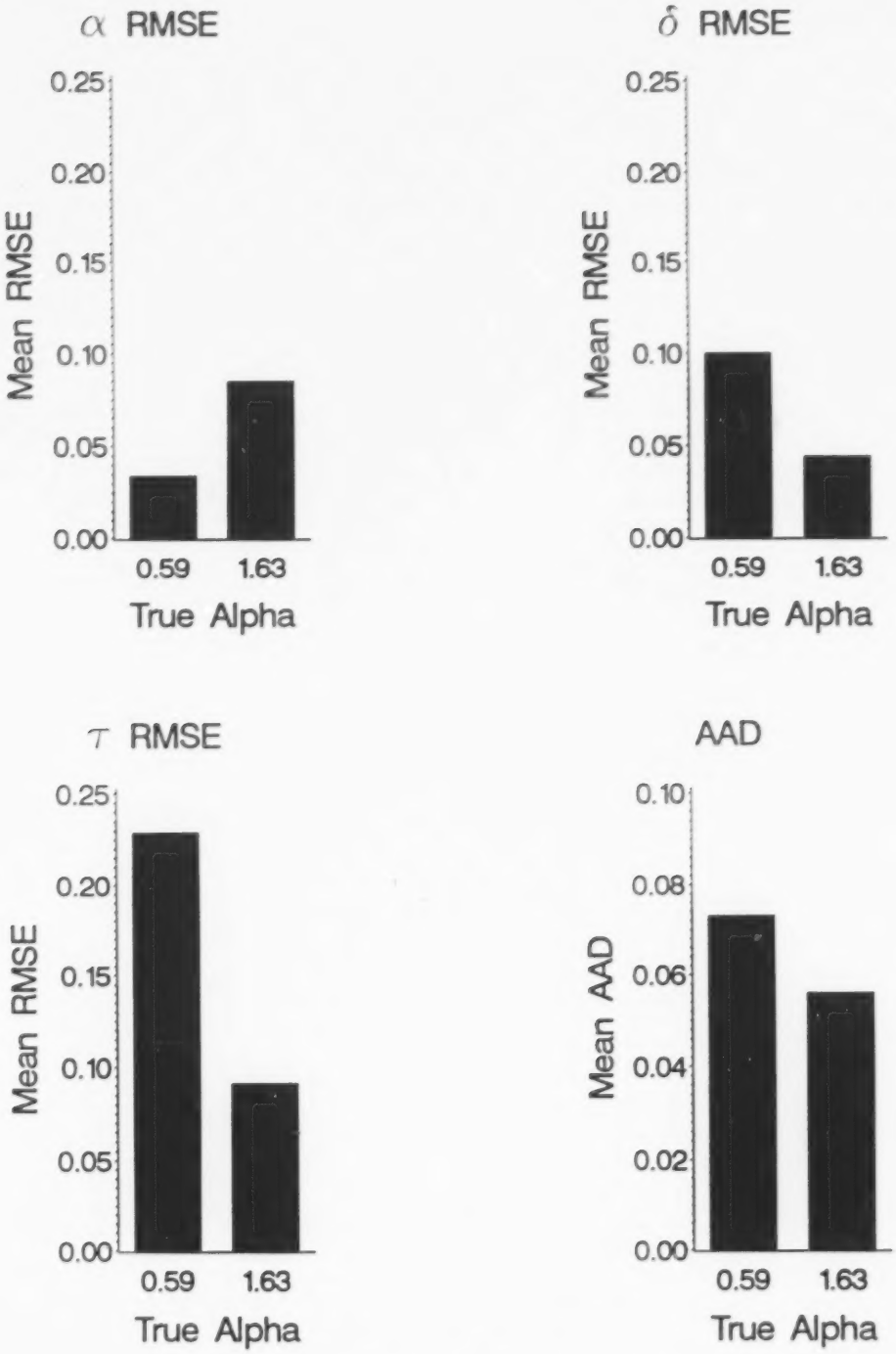

Figure 3. The effect of true α_i on average RMSE values for α_i (upper left panel), δ_i (upper right panel), and τ_{ik} (lower left panel) parameter estimates. The corresponding effect on mean AAD values is shown in the lower right panel.

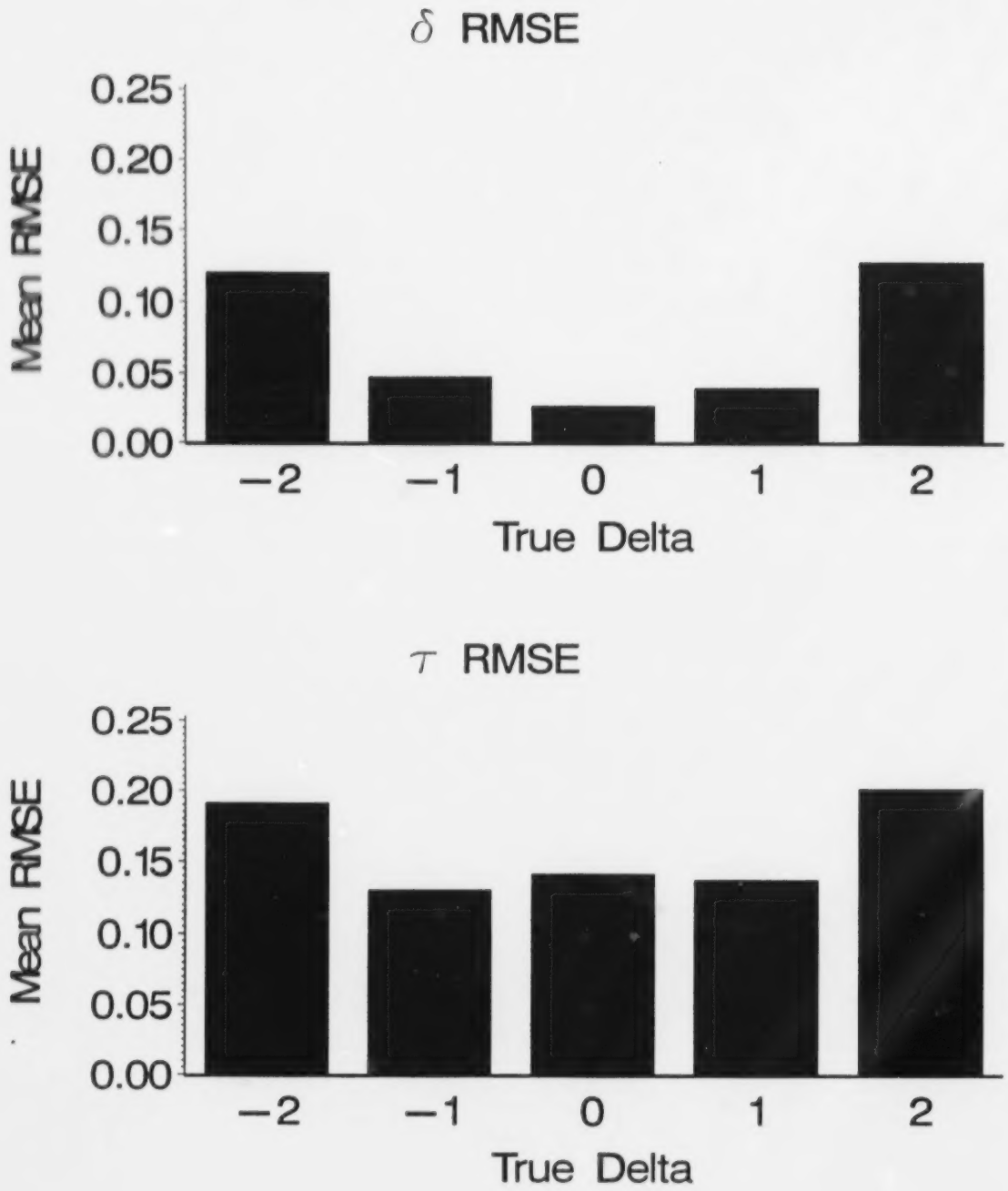

Figure 4. The effect of true δ_i on the mean RMSE values for δ_i (upper panel) and τ_a (lower panel) parameter estimates.

Discussion

The results of this research suggest that the inaccuracy of a given GGUM item parameter estimate will vary depending on the true value of the item parameter in question (same parameter effect) and on the true values of other item parameters in the model (cross parameter effects). However, the degree of both types of parameter estimation inaccuracy will typically be modest. These results also imply that both types of parameter estimation inaccuracy will have negligible effects on the precision of estimated item response functions. Thus, while interrelationships among parameters will reliably affect their estimation and the ultimate representation of the response function, these effects will generally be small in magnitude and will have little practical consequences as long as 1) the data truly conform to the GGUM, 2) a sufficiently large sample of respondents (i.e., $N \geq 750$) is used in the MML estimation procedure, and 3) the distributions of true parameters are similar to those used here. Obviously, one must be cognizant of the limited generality of a single simulation study, but these results do suggest that MML estimation of GGUM item parameters and corresponding item response functions are fairly robust to both same parameter and cross parameter effects.

The ability to accurately estimate GGUM item parameters is a prerequisite for the general application of the model in applied measurement situations. These results provide further support that all of the GGUM item parameters are estimable given sufficient sample sizes. Moreover, given accurate parameter estimates and a well-fitting model, the GGUM should provide the same benefits of other item response theory models such as 1) item parameter invariance, 2) person parameter invariance, and 3) the ability to estimate the precision of a single individual's attitude estimate. These benefits set the stage for other measurement advantages associated with item response theory models such as item banking and computer adaptive testing. Therefore, the GGUM should prove to be a useful tool for large scale assessment of attitudes and preferences in educational settings.

References

Andrich, D. (1996). A general hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49*, 347-365.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Coombs, C. H. (1964). *A theory of data*. New York: John Wiley & Sons.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 159-173.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Roberts, J. S. (1995). Item response theory approaches to attitude measurement. (Doctoral dissertation, University of South Carolina, Columbia). *Dissertation Abstracts International, 56*, 7089B.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1998). *The generalized graded unfolding model: A general parametric item response model for unfolding graded responses* (RR-98-32). Princeton, N.J.: Educational Testing Service.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1999). A general model for unfolding unidimensional polytomous responses using item response theory. *Applied Psychological Measurement*, in press.

Roberts, J. S., & Laughlin, J. E. (1996a). *The graded unfolding model: A unidimensional item response model for unfolding graded responses* (RR-96-16). Princeton, NJ: Educational Testing Service.

Roberts, J. S. & Laughlin, J. E. (1996b). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*, 231-255.

Roberts, J. S., Laughlin, J. E., & Wedell, D. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59*, 211-233.

van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis is often the incorrect model for analyzing bipolar concepts, and what model can be used instead. *Applied Psychological Measurement, 18*, 97-110.